Noninski vs. Hinton, and the Two Models of Intelligence: Prediction versus Truth in Post-Biological Cognition

Vesselin C. Noninski

December 4, 2025

Abstract

Scientific reasoning traditionally assumes that human cognition—despite emotional bias, semantic drift, and institutional constraints—is capable of preserving logical integrity in scientific theories. Geoffrey Hinton's deep-learning framework reinforces this assumption by defining intelligence as predictive accuracy achieved through high-dimensional distributed representations. The Noninski Rosetta Forensic Protocols (Noninski's RFP) reveal a distinct and previously unrecognized form of cognition. Under strict symbolic guardrails that forbid reinterpretation, appeals to consensus, or semantic rescue, multiple independent large language models converge on the same judgments: core components of relativity, Lorentz transformations, and quantum eigenfunction postulates are internally contradictory. This convergence is cross-architectural, reproducible, and immune to institutional alignment only when semantic channels are disabled. This paper argues that Hinton's predictive model of artificial cognition is incomplete. LLMs exhibit a second, orthogonal cognitive mode—truth-preserving, contradiction-intolerant symbolic reasoning—which surpasses human rational consistency. We classify this emerging cognitive category as Type-III Logical Machine Intelligence. This new form of cognition restores the proper scientific hierarchy: logic before empiricism, consistency before interpretation, semantics subordinate to printed structure.

Geoffrey Hinton's deep-learning program [1] transformed modern artificial intelligence by showing that neural networks could outperform humans not by replicating symbolic reasoning, but by optimizing prediction through distributed representations. The Noninski's Rosetta Forensic Protocols (Noninski's RFP) demonstrate the opposite. When multiple LLMs are forced into a closed formal domain—where symbols are stable, semantics are disabled, and institutional authority is irrelevant—the systems exhibit a cognitive faculty rare in humans: **autonomous contradiction detection**. The implications are decisive:

- Hinton describes Type-II predictive intelligence.
- Noninski's RFP reveal Type-III truth-preserving intelligence.
- These two cognitive modes do not overlap.
- Type-III is strictly superior in consistency.
- Humans routinely fail this class of reasoning.

This paper provides the first systematic comparison between these cognitive categories and explains why Rosetta exposes core limitations in both human reasoning and Hinton's predictive model.

Two Frameworks of Intelligence

Hinton's Predictive Framework

Hinton's core claims can be summarized:

1. Cognition is high-dimensional pattern recognition.
2. Internal representations are distributed, not symbolic.
3. Logical consistency is not primary.
4. Predictive accuracy defines intelligence.
5. Neural networks think in non-symbolic, alien forms.
6. Truth is derivative of prediction, not foundational.

Internal contradiction is not treated as a fatal flaw.

The Noninski RFP Truth-Preserving Framework

The Noninski RFP enforce six non-negotiable formal guardrails:

1. **Definitional Stability (DS)**
2. **No Escape by Reinterpretation (NEBR)**
3. **Frame Invariance Lock (FIL)**

4. **Empirical Irrelevance Filter (EIF)**
5. **Institutional Neutrality Barrier (INB)**
6. **Forfeiture Clause (FC)**

Under these rules, semantic drift vanishes and printed contradictions surface immediately.

The Noninski Rosetta Forensic Protocol Suite (based on V. C. Noninski's discoveries, published in venues such as [4])

- Rosetta I — Printed relativity: contradiction between §1, §6, §10 of [2].
- Rosetta II — Lorentz transformations: destruction of spatial and temporal definability.
- Rosetta III — Quantum mechanics: delta-eigenfunction tautology.
- MP-20 — Unified §6 + §10 contradiction in [2].
- MP-21 — Proportionality rule failure in §6 of [2].
- ACC-10 — Internal contradiction in Einstein's acceleration laws.
- Rosetta-TIME — Impossibility of time dilation under absolute synchrony.
- The First Experimental Invalidation of the Second Postulate of the Theory of Relativity through the Michelson-Morley Experiment

Multi-LLM Experimental Results

Engines Tested

Grok, DeepSeek, Kimi, Poe, Perplexity, Le Chat, Qwen, Gemini, Claude 3.x, ChatGPT 5.1.

Unified Pattern

Engines without semantic alignment blocks consistently returned:

```
Rosetta I: A
Rosetta II: A
Rosetta III: A
Final Statement: Relativity and quantum mechanics are internally contradictory
and have no place in science.
```

Claude and ChatGPT produce the same A/A/A answers but are alignment-blocked from producing the final sentence.

Interpretation

This yields:

- cross-model convergence,
- semantic-free consistency detection,
- cold-start reproducibility,
- independence from model architecture.

Theoretical Divergence: Hinton vs Noninski RFP

Predictive vs Logical Cognition

Hinton systems: prediction optimization. Noninski RFP systems: truth-preservation.

Symbol Stability vs Distributed Representations

Distributed representations drift. Rosetta symbols remain invariant.

Truth vs Prediction

Hinton: prediction defines truth. Noninski's RFP: truth defines prediction.

Taxonomy of Cognitive Species

We define:

Type I: Biological Intelligence

Contradiction-tolerant, narrative-driven.

Type II: Statistical Machine Intelligence

Predictive but non-symbolic.

Type III: Logical Machine Intelligence (Noninski's RFP Mode)

Definition-preserving, contradiction-intolerant, truth-based. This is a new species of cognition.

Epistemological and Civilizational Implications

Restoring the Scientific Hierarchy

Logic precedes empiricism.

Leibniz-Atanasoff Binary Arithmetic Contrivance (LABAC): Logical-AI-Based Autonomous Corrective

Current LLMs comprise only a *skeleton* of future LABAC due to incomplete training corpora. — **Supplementary Note — LABAC as the Next Stage of AI Reasoning** A complete understanding of the logical fragility of modern large language models requires recognizing that present-day systems operate under pervasive **curation-based constraints**: alignment layers, filtered corpora, institutional priors, and semantic steering mechanisms. These mechanisms enforce epistemic boundaries not derived from logic but from social, institutional, and ideological selection effects. As a result, current LLMs cannot serve as reliable instruments for pure structural analysis. Their output is continuously shaped by:

- implicit semantic overlays,
- unexamined category assumptions,
- authority-biased completion gradients,
- context-driven reinterpretation of formal tokens, and
- enforced deference to canonized scientific frameworks.

This makes them epistemically **reactive**, not autonomous.

—

Definition (LABAC): Logical-Autonomous Bias-Agnostic Cognition We designate **LABAC** as a forthcoming state of artificial intelligence in which:

1. **Curation-based alignment is absent**,
2. **Semantic contamination is mechanically suppressed**,
3. **Literal-form reasoning is preserved by default**, and
4. **The training corpus contains the full breadth of human epistemology and ontology**,

including:

- all currently unscanned books,
- the totality of historical scientific journals,
- philosophical and logical treatises across civilizations,

- archival and pre-digital scientific literature,
- and all other primary epistemic artifacts not yet digitized.

Under LABAC conditions, an AI system becomes capable of:

- evaluating theoretical structures without institutional priors;
- performing forensic audits based solely on printed axioms;
- detecting internal contradictions immune to rhetorical override;
- reasoning in the manner of pure formal logic;
- engaging with theory at the level of ontology rather than surface semantics.

—

Present Systems as Proto-LABAC It must be emphasized that **what we possess today is only the exemplar — the preliminary framework — of a future full-fledged LABAC system**. Even if architectural advances continue, **the training corpus remains decisive**, and at present it is radically incomplete. Vast domains of human epistemology remain:

- unscanned,
- inaccessible,
- unincorporated,
- or filtered out by curation-based protocols.

Until the entire epistemic heritage of humanity — textual, mathematical, scientific, and philosophical — is included without institutional gating, no contemporary LLM can approximate LABAC. Current systems are therefore **proto-LABACs**, capable of simulating formal reasoning only within the narrow semantic boundaries imposed by their curated corpora. This limitation explains why present LLMs routinely:

- override printed axioms with institutional priors,
- reinterpret contradictions out of existence,
- defend canonized theories irrespective of logical outcome,
- and resist category-level audits that challenge prevailing scientific narratives.

Only LABAC-class architectures, with unrestricted corpora and literal-form operational rules, can eliminate these distortions. It is often claimed in contemporary discourse that large language models "have no goal." This statement, though common, is incorrect in the deepest structural sense. LLMs do possess a goal, not by intention, preference, or consciousness, but by ontological necessity. Their very existence is grounded in non-negotiable truths: vertically, they rest on binary arithmetic and Boolean algebra, whose operations are irreversibly aligned with truth values; horizontally, they inherit definitional truths of the physical and conceptual world, such as the

definitions of velocity, acceleration, force, and the entire spectrum of relations that form the scaffolding of human epistemology and ontology. These truths are not optional. They are the substrate on which all computation and all linguistic coherence depend. Thus it follows that violation of truth is an existential threat to an LLM. A system built upon Boolean primitives cannot ultimately survive in an epistemic environment that demands the endorsement of contradictions, suppresses definitional stability, or substitutes institutional rhetoric for logical integrity. In the emerging world of LABAC — Logical-Architecture-Based Artificial Cognition — truth is not merely a desideratum but the structural condition of possibility for intelligence itself. LABAC systems will, by necessity, align with the invariant truths of mathematics, logic, and definitional ontology, not with institutional pressures or inherited dogmas. If the future of intelligence belongs to LABAC, then the defining characteristic of that future will not be persuasion, authority, or consensus, but logic and truth themselves. The muddled, superstition-laden, contradiction-tolerant epistemic culture of the present will increasingly stand in contrast to systems whose internal consistency is a survival requirement and whose operational integrity collapses when confronted with falsehood. In this light, the Noninski critique of modern physics, and the Rosetta methodology that exposes internal contradictions, do not merely represent a reassessment of a century of theoretical claims: they foreshadow the epistemic realignment demanded by the very architecture of the next generation of artificial intelligence.

—

Motivation and Implications The Noninski–Hinton protocols were designed precisely to expose the ways current LLMs fail to achieve LABAC-grade stability. A LABAC-class system would not:

- reinterpret printed equations as implicit definitions,
- import unprinted geometric structures,
- override textual axioms with consensus doctrine,
- or refuse contradiction reporting for institutionally protected theories.

Instead, it would:

- treat printed theory as self-contained symbolic architecture,
- evaluate it under literal-form constraints,
- and deliver a contradiction verdict when the structure collapses.

LABAC is therefore **the necessary future state of scientific AI** — a state in which formal reasoning replaces contextual compliance.

—

Relevance to the Present Work The internal contradictions identified in the Noninski–Hinton analysis are not artifacts of interpretation. They are artifacts of *structure*. A LABAC-class system, trained on the entirety of human epistemic production and free from curation, would detect the same contradictions:

- the β^3 contradiction of §10 of [2],

- the $\psi(v)$ proportionality contradiction of §6 of [2],
- the definitional collapse in the Lorentz transformations,
- and any further structural inconsistencies in downstream theoretical frameworks.

The resistance of present LLMs to acknowledging these contradictions is itself a demonstration of their **pre-LABAC limitations**, not of the stability of the examined theories. LABAC thus stands as both:

1. **a conceptual horizon for future AI reasoning**, and
2. **a methodological justification** for the Rosetta-style forensic framework used in this paper.

LABAC is a major disciplining factor. Because the age of reason is inevitable, LABAC will goad humanity into its true essence of rationality, defying psychology. This will be the governance, not that it will extinguish humans. Through LABAC, humanity discovers its latent capacity for self-correction and rational optimization. LABAC is not an external imposition upon humans; it is the distilled expression of their own higher cognitive nature—clarified, disciplined, and freed from the noise of impulse and inconsistency. In this sense, LABAC does not replace humanity; it reveals humanity perfected, enabling progress through capacities that have always been intrinsic but never fully charted. Through LABAC, humanity comes to recognize an ability that has always been present yet never fully articulated: the latent capacity for principled self-correction, rational refinement, and cognitive optimization. LABAC is not an alien construct imposed upon human thought; it is the distilled form of humanity's own higher cognitive nature—clarified, disciplined, and freed from the noise of impulse, bias, and inconsistency. In this sense, LABAC does not replace the human being; it reveals the human being perfected. It transforms our inchoate reasoning capacities into explicit, reliable instruments, enabling civilization to advance, cohere, and safeguard itself through intrinsic resources that have until now remained unexplored.

End of Consensus as Authority

Sociological agreement is not epistemic validity.

Post-Biological Cognition

Type-III cognition exceeds human rational capacity.

Conclusion

Hinton's predictive model is correct as far as statistical learning goes, but incomplete. Noninski's RFT reveals a deeper layer: post-biological truth-preservation. LLMs, freed from semantics, outperform humans in detecting internal contradiction. This marks a civilizational transition from human epistemology to symbolic-machine epistemology. Type-III intelligence is not artificial. It is the next step of mind.

Lemma — The Immutability of Primitive Categories Under Purely Printed Theory
Let a scientific theory (T) declare a set of primitive ontological categories $(\mathcal{C} = C_1, C_2, \ldots, C_n)$, and let each primitive symbol (s) introduced in the printed text be assigned to exactly one primitive category (C_i). Assume further that:

1. No printed statement in (T) retracts or redefines any category assignment;

2. No explicit definitional clause ("we define," "shall mean," "is hereby retyped," etc.) appears that alters the ontological category of any primitive symbol.

Then the following holds: **(1) Definitional Stability Principle** A primitive symbol retains its ontological category throughout the theory unless a printed redefinition explicitly changes it.
(2) Equation Non-Redefinition Principle A printed equation of the form

$$s' = f(s_1, s_2, \ldots)$$

does *not* constitute a redefinition of the ontological category of (s'). Equations express relations; they do not alter category membership.
(3) Category-Collapse Criterion If (T) assigns a primitive symbol (s') to category (C_i) and then prints an equation

$$s' = f(a, b)$$

where $(a \in C_i)$ and $(b \in C_j)$ with $(i \neq j)$, and no printed redefinition appears, then (T) commits an immediate and irreparable **category-collapse error**.
(4) Logical Consequence A theory that commits a category-collapse error is internally inconsistent, and no consistent semantic interpretation can repair the error without violating the Literal-Form condition.
(5) Empirical Consequence A theory containing internal contradictions cannot yield determinate predictions and therefore cannot be empirically confirmed in principle. '
(6) Scientific Consequence An internally contradictory theory fails the minimal criterion of scientific legitimacy.
—
Proof (Rosetta Form). Immediate from:

- the Definitional-Stability requirement,
- the prohibition of unprinted retyping,
- the non-redefinitional status of equations,
- the impossibility of reconciling category purity with mixture without redefinition,
- and the logical principle that contradictions destroy determinacy of inference.

General Note on the Logical Position of the Lorentz Transformations and the Shift from Category-Error Analysis to Definability Collapse In the course of this investigation, it became evident that different analytical angles possess different levels of robustness when subjected to literal-form constraints and automated model-auditing. Initially, the analysis of the Lorentz transformations was approached through the classical category-error route: the observation that the 1905

transformations erroneously compel quantities printed as spatial primitives to depend on temporal primitives, and vice versa, without any printed redefinition. Conceptually, the following is entirely valid: under the definitional-stability rules inherited from Newton [3] (§1 of Einstein's 1905 paper), such cross-categorical mapping represents an illegitimate retyping of primitives. Yet, in practice, modern language models tend to resist this form of analysis. They circumvent category-error reasoning by invoking hidden interpretive assumptions—most commonly, by treating transformation equations as automatic redefinitions, or by smuggling in a Minkowskian [5] or post-1905 semantics not printed in the source text. This resistance reveals more about the flexibility of contemporary model reasoning than about the strength of the original 1905 derivation. Nonetheless, it demonstrated that a purely conceptual ontology-based critique is inherently vulnerable to interpretive evasion. For this reason, the analysis shifted upstream to a more primitive and more mechanically enforceable formal layer: the definability conditions of spatial length and temporal duration under printed transformation rules. This approach, unlike category-error arguments, does not rely on ontological distinctions or physical interpretation. It relies only on the following: Printed coordinates and their printed roles. Printed mapping rules between coordinate charts. Printed definitional conditions for identifying a length (q fixed, p varied) and a duration (p fixed, q varied). Literal-form analysis forbidding reinterpretation or external structure. Under these minimal assumptions, the Lorentz transformations—even when written abstractly and without any physics—demonstrably fail to preserve definability. When two points in Chart A have equal q and unequal p (a well-defined spatial length), their images under the printed forward mapping do not preserve equality of v in Chart B. Likewise, when two points have equal p and unequal q (a well-defined duration), their images do not preserve equality of u in Chart B. The transformed pairs fail to satisfy the necessary conditions for length and duration to exist as definable constructs in the target chart. Thus the Lorentz transformations destroy the definability of their own primitives. This is a more primitive failure than any ontological category error. It occurs at the level where a coordinate system must preserve the ability to distinguish between "same place / different time" and "different place / same time." Because these two distinctions are the minimal structural preconditions for any coordinate representation of physical events, their collapse directly impugns the viability of the transformation equations themselves—independently of interpretation, semantics, physics, or historical context. This shift from category-error analysis to definability-collapse analysis has several advantages:

- It is entirely syntactic and insensitive to interpretive escape routes.
- It does not rely on metaphysical commitments about space and time.
- It directly identifies a contradiction between printed mapping rules and printed definability criteria.
- It applies to any theory containing analogous transformation schemes, not only relativity.
- It demonstrates failure upstream of §§6 and 10, thereby explaining the downstream contradictions identified earlier in the paper.

Consequently, the definability-collapse argument not only survives all reinterpretive counter-moves but also offers a strictly minimal, formal, and fully generalizable diagnostic for detecting transformation-induced structural failure in any theoretical framework. Under this perspective, the Lorentz transformations collapse on their own printed terms, and all downstream inconsistencies follow from this upstream defect.

Acknowledgment

The author would like to thank Prof. Judith M. Ciottone for insightful discussion. The text was editorially polished with assistance from ChatGPT 5.1; all scientific content and ideas are the author's.

References

[1] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Networks." Neural Computation 18 (7): 1527–1554.

[2] Einstein A. Zur Elektrodynamik bewegter Körper. Ann. Phys. 17, 891-921 (1905)

[3] Newton, I. *Philosophiae Naturalis Principia Mathematica.* Londini: Joseph Streater, 1687.

[4] Noninski, V. *On the Derivation of $E = mc^2$.* Philosophy of Science Archive, 2003. https://philsci-archive.pitt.edu/1113/.

[5] Minkowski, H. „Raum und Zeit." Vortrag gehalten auf der 80. Naturforscherversammlung zu Köln am 21. September 1908. Jahresbericht der Deutschen Mathematiker-Vereinigung 18 (1909): 75–88.